REPORT DOCUMENTATION PAGE

OMB No. 0704-0188

AD-A263 704

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
1 May 93	Final Report 1Sep89-28Feb93

4. TITLE AND SUBTITLE

Empirical Quantile Function Nonparametric Integrated Analyses

5. FUNDING NUMBERS

DAAL03-89-G-0098

6. AUTHOR(S)

W.D. Kaigh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Department of Mathematical Sciences
University of Texas at El Paso
El Paso, Texas 79968

PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ARO 27159.9-MA-SAH

11. SUPPLEMENTARY NOTES

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

93-09738

93 5 05 031

15pg

13. ABSTRACT *(Maximum 200 words)*

This three-year statistics research project addressed theoretical development, efficiency evaluation, and desktop computer implementation of an integrated nonparametric statistical data analysis approach based on Fourier analytic techniques applied to the empirical quantile function (EQF). New EQF statistical procedures were developed for data smoothing and reduction methods, nonparametric estimation of various functionals associated with the quantile function, composite goodness-of-fit tests for uniform and exponential models with applications to related stochastic processes, and nonparametric analysis of variance rank procedures for analysis of independent samples.

These EQF methods exploit discrete Hahn polynomial orthogonal polynomial component representations to produce statistics for significance testing and interval estimation for both omnibus and directional alternative models. Performance evaluations of the proposed techniques included theoretical and Monte Carlo efficiency comparisons as well as numerical applications to challenging data sets. Interactive desktop computer and graphics routines utilizing symbolic programming languages were developed to facilitate implementation of EQF data analysis techniques by statistical users.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Nonparametric Statistics, Quantile Function, Goodness of Fit, Components	13
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD	- See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE	- See authorities.
NASA	- See Handbook NHB 2200.2.
NTIS	- Leave blank.

Block 12b. Distribution Code.

DOD	- Leave blank.
DOE	- Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA	- Leave blank.
NTIS	- Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

EMPIRICAL QUANTILE FUNCTION NONPARAMETRIC INTEGRATED ANALYSES

Final Report

W. D. Kaigh

May 1, 1993

U. S. Army Research Office

Grant Number DAAL03-89-G-0098

University of Texas at El Paso

THE VIEWS, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION

EMPIRICAL QUANTILE FUNCTION NONPARAMETRIC INTEGRATED ANALYSES

A. Statement of the Problem Studied

The primary objective of this theoretical statistics research investigation was to extend empirical quantile function (EQF) methods as alternatives to traditional empirical distribution function (EDF) procedures in various nonparametric statistical applications. Due to the minimal population assumptions required for their application and validity, the thoroughly documented efficacy of nonparametric procedures is now widely accepted by statistical data analysts.

With a seminal statistical approach reminiscent of classical Fourier analysis, Durbin and Knott (1972) introduced component decompositions of conventional EDF quadratic goodness-of-fit statistics to detect population differences not identifiable from a statistically significant omnibus result. In particular, they endorsed generally examining individual component statistics to augment standard statistical analyses. Analogous to principal components analysis in multivariate statistics, decomposition of quadratic test statistics into squares of standardized individual components (uncorrelated with zero means and unit variances under a null hypothesis) has since received considerable attention in the statistical literature. An excellent overview with substantial background material appears in Shorack and Wellner (1986, chapter 5).

In another seminal work, Parzen (1979) advocated alternative statistical approaches based on the population quantile function and its sample analogue the EQF. To extend the components approach to EQF estimation and testing procedures for various statistical applications, specific research objectives for this project encompassed (i) development of point and interval estimation techniques for various functionals associated with the quantile function; (ii) development of quadratic goodness-of-fit EQF tests with directional spacings component decompositions for distributional assessment; (iii) development of distribution-free omnibus quadratic EQF tests with directional rank spacings component decompositions for nonparametric analysis of variance procedures.

As well as to seek and explore related nonparametric statistical EQF applications, the general intent of this research project was to

- Develop distribution theory for individual EQF components and aggregate quadratic statistics

- Evaluate relative performances of individual EQF components and aggregate statistics

- Obtain standard error estimates and interval estimation procedures for individual EQF statistics

- Provide meaningful statistical interpretations of EQF components

- Develop desktop computer EQF numerical and graphical routines for statistical users.

B. Summary of Most Important Results

Several new EQF statistical techniques were developed for nonparametric population quantile estimation, data smoothing and reduction, goodness-of-fit tests for uniform and exponential model distributional assessment, and the two-sample nonparametric problem. These fundamental EQF results provide a firm basis for subsequent modification and generalization to related nonparametric statistical applications.

Specific details of these EQF integrated procedures appear in Kaigh and Cheng (1991), Kaigh (1992), Kaigh (1993), Kaigh and Hosch (1993), Kaigh and Sorto (1993), Kaigh and White (1993), Sorto (1992), and White (1992). For convenience of those reviewing this report, titles and technical abstracts for these manuscripts appear in Section C.

Within broad categories, brief outlines of the major statistical results developed under the project are as follows:

- EQF data reduction and smoothing (generalized order statistics obtained from Kaigh-Lachenbruch (KL) data reduction and Kaigh-Cheng (KC) isotonic data smoothing linear transformations, Lorenz partial-order majorization inequalities);

- EQF quantile functional estimation (generalized point and interval estimation of population quantiles based on L-statistics obtained from KL and KC generalized order statistics, continuous functional estimators derived from Bernstein polynomial extension of KL and KC quantile function and Lorenz curve point estimators);

- EQF uniformity criteria (quadratic goodness-of-fit tests with L-statistic component decompositions based on sample spacings from the KL and KC generalized order statistics);

- EQF exponentiality criteria (total time on test function quadratic goodness-of-fit criteria with L-statistic component decompositions from KL and KC generalized order statistics);

☛ EQF rank statistic criteria (two-sample quadratic tests with rank spacings component decompositions).

Traditional EDF techniques focus on the sample moments, whereas EQF methods employ the sample order statistics and spacings. Although analytically related to previous orthogonal components work with EDF statistics, these integrated statistical analysis techniques emphasize quantile functional estimation and component decompositions of related EQF quadratic tests. In spirit as well as in similarity of analytic methods, the new EQF procedures adhere to the comprehensive data analysis approach espoused by Rayner and Best (1989).

For conceptual understanding and classification, slightly expanded summary descriptions of the major research accomplished under the project are given below.

Quantile Functional Estimation

Extending results in Kaigh and Cheng (1991), point estimators of Legendre polynomial Fourier coefficients for the quantile density function and total time on test function derivative are given in Kaigh (1992) and Kaigh and White (1993), respectively. Incorporating these latter results with those in Kaigh and Sorto (1993), these data smoothing approaches blend KC smoothing transformations with Bernstein and related polynomial approximations of the EQF. Related quantile functional estima.)n work appears in the applications paper by Kaigh and Hosch (1993).

The KL and KC subsampling quantile estimators with jackknife and bootstrap standard errors produce effective alternative point and interval estimation procedures for specific values of the population quantile function, say, quartiles or deciles. Most recently, Kaigh and Sorto (1993) obtained isotonic and majorization results for the corresponding Lorenz curves and then employed Bernstein polynomial type approximation schemes with the subsampling quantile estimators to produce continuous functional estimators of the EQF and population Lorenz curve for assumed continuous data.

EQF Goodness-of-Fit Component Decompositions

Related to several recent statistical applications which parallel classical Fourier analysis, EQF procedures employ discrete Hahn polynomial orthonormal vectors to produce quadratic omnibus test statistics. Orthogonal component representations show that EQF quadratic criteria are sums of squares of individual EQF component statistics. Easily interpreted and now readily computable, these individual components permit assessment of specific directional departures from an assumed statistical model.

Diagonalization of appropriate covariance matrices is the fundamental mathematical problem implicit in component decompositions of quadratic statistics, but corresponding eigenvalue

problems are seldom tractable and usually require numerical solution. Despite complex covariance structures typically associated with order statistics, this investigation exploited continuous Legendre and discrete orthogonal Hahn polynomial vector diagonalization results for each of the EQF applications listed previously in Section A.

These general diagonalization and component aggregate results reproduce the Greenwood spacings statistic, the Dixon two-sample rank spacings statistic, and EQF analogues of the conventional EDF Anderson-Darling and Neyman smooth tests as special cases. In addition, several new quadratic test criteria and individual component statistics, emerged quite naturally from the general theoretical approach. Moreover, because the corresponding matrices have Hahn polynomial orthonormal vector spectral representations, these individual component statistics are preserved by the KL reduction and KC smoothing transformations.

EQF Extensions

Although peripheral to the statistical focus and objectives of the research project, goodness-of-fit applications to stochastic processes and time series analysis as well as unexpected mathematical results emerged indirectly as by-products from the theoretical statistical investigations. Specifically, it appears that new mathematical derivations and results with potentially general ramifications were obtained in the applied mathematics disciplines of linear algebra, orthogonal functions, approximation theory, and isotonic and majorization methods. Further development and preparation of these results for publication in the applied mathematics literature is currently underway.

Statistical Computation

Extensive computational effort was expended to facilitate desktop computer utilization and implementation of EQF data analysis techniques by statistical users. Although many of the one and two-sample EQF procedures can be performed with only desktop computer spreadsheet software, most statistical users are unlikely to adopt new techniques requiring more than minimal programming effort. Consequently, simple computational methods suitable for use in an interactive computing environment were deemed essential project objectives.

As documented in Sorto (1992), White (1992), and Kaigh and Hosch (1993), computational and graphical features available with the symbolic mathematical programming languages MATHEMATICA (or MAPLE) have virtually eliminated any statistical computing obstacles to preclude usage of EQF data analysis procedures. In many cases, previously developed complicated Fortran computational routines requiring over a page of computer code were replaced by simple one-line MATHEMATICA commands employing supplied functions to yield any specified numerical precision.

Finally, it should be noted that the intent of the EQF integrated approach is not to supplant conventional EDF procedures, but instead to augment these proven techniques with parallel EQF analyses. In fact, the major original statistical contribution of this effort may be the development of nonparametric statistical procedures which incorporate both EDF and EQF sample information.

C. Publications and Reports

As previously indicated, several technical research manuscripts were prepared during the project. With corresponding literature citations in order of most recent appearance, titles and abstracts for these manuscripts are listed below.

Total Time on Test Function Orthogonal Components and Tests of Exponentiality

W. D. Kaigh and Alexander K. White (1993)
Proceedings of the 37th Conference on the Design of Experiments (In press)

Elementary mathematical development reminiscent of Fourier analysis applied to the sample total time on test function (TTT) yields scale-free orthogonal components analogous to empirical quantile function (EQF) component L-statistics utilized by Kaigh (1992) for assessing one-sample uniformity. The TTT components are linear combinations of normalized spacings with Hahn polynomial vector weight functions to provide directional criteria for assessment of departures from exponentiality. In particular, the first TTT component is equivalent to the cumulative total time on test statistic and Gini statistic studied in Gail and Gastwirth (1978a). Analogous to the quadratic smooth tests for exponentiality proposed by Rayner and Best (1986, 1989), aggregates of squared TTT components yield component decompositions of a discrete Anderson-Darling type statistic and the squared coefficient of variation. Monte Carlo results indicate adequacy of asymptotics for small samples and empirical power comparisons show that TTT component exponentiality criteria are quite competitive for various alternative models including those with "bathtub" hazard rates.

Subsampling Quantile Estimator Majorization Inequalities

W. D. Kaigh and Maria Alejandra Sorto (1993)
Statistics and Probability Letters (In press)

A Lorenz partial order majorization inequality is obtained for the Kaigh-Lachenbruch (KL), Harrell-Davis (HD), and Kaigh-Cheng (KC) subsampling quantile estimators developed in Kaigh

and Cheng (1991a,b) for n numerical data values. Incorporating a subsampling data-reducing parameter k, the general result for sample Lorenz curves shows that $L_{KC} \leq L_{HD} \leq L_{KL}$ pointwise for any $k=1,\ldots,n$. Specific application in the case $k=n$ demonstrates that KC generalized order statistics provide the "smoothest" values and that both HD and KC generalized order statistics are "smoother" than the usual set of n order statistics. It follows from Schur convexity and the majorization theorem that, among the three types of subsampling designs, conventional statistical measures of dispersion such as the standard deviation, mean absolute deviation, Gini mean difference, coefficient of variation, and (negative) entropy are always least for the KC ordered values. For statistical data from a continuous population, utilizing the three different sets of order statistics as discrete input for Bernstein polynomial approximation schemes yields continuous quantile function estimators which are also Lorenz ordered according to their respective inputs.

Distribution-Free Two-Sample Tests Based on Rank Spacings

W. D. Kaigh (1993)
Journal of the American Statistical Association (In press)

For the nonparametric two-sample problem, rank spacings of one sample are the positive integer distances between combined-sample ordered ranks from that sample. Phrased in terms of the algebraically equivalent spacings frequencies treated by other authors, rank spacings are obtained by addition of constant value one to the number of observations from the second sample separating consecutive ordered values of the first sample. Under the two-sample null hypothesis, rank spacings are exchangeable random variables with constant means.

Elementary mathematical development reminiscent of Fourier analysis yields distribution-free orthogonal components analogous to L-statistics utilized by Kaigh (1992) for assessing one-sample uniformity. Individual rank spacings components are linear combinations of ordered ranks with Hahn polynomial vector weight functions to provide directional criteria for assessment of between-sample differences. The first four rank spacings components provide nonparametric measures of location, scale, skewness, and kurtosis. Classified as nonlinear rank exceedance statistics in the two-sample nonparametric context, the asymptotically normal rank spacings components are related to linear rank statistic components of the two-sample Anderson-Darling statistic obtained by Pettitt (1976). Aggregates of squared rank spacings components yield a component decomposition of the Dixon (1940) statistic and provide omnibus chi-squared statistics analogous to those in Pettitt (1976) and Boos (1986). Monte Carlo results indicate adequacy of asymptotics for small samples and empirical power comparisons show that rank spacings analysis

is quite competitive with conventional two-sample procedures.

Because the implicit fundamental mathematical problem is assessment of a simple random sampling assumption from a finite population, two-sample rank spacings components lead naturally to a variety of one-sample goodness-of-fit procedures based on the empirical distribution function (EDF) and empirical quantile function (EQF). These include the Greenwood (1946) spacings statistic, the Anderson-Darling statistic, and the Neyman smooth tests as well as their EQF analogues developed in Kaigh (1992).

Quartile Estimation Techniques: Explanation and Application to Data Analysis in Psychology

W. D. Kaigh and Harmon M. Hosch (1993)

Quartile estimators, bootstrap standard errors, and interval estimation and hypothesis testing methods for one and two-sample nonparametric statistical data description and analysis are presented. The proposed procedures are based on conventional sample quantiles as well as recently developed alternative methods of quantile estimation. The quartile estimation techniques are applied for statistical analysis and interpretation of treatment contrasts with three experimental data sets.

EDF and EQF Orthogonal Component Decompositions and Tests of Uniformity

W. D. Kaigh (1992)
Nonparametric Statistics, 1, 313-334.

Orthogonal expansions of a probability density function and the corresponding quantile density function are employed to motivate new as well as existing omnibus quadratic tests for uniformity. Utilizing Legendre polynomial component decompositions, the proposed goodness-of-fit criteria are based on Fourier analytic techniques applied to the empirical distribution function (EDF) and the empirical quantile function (EQF).

Individual EDF components involve the continuous Legendre polynomials and component aggregates provide the Neyman smooth statistics. The EQF components are spacings statistics with Hahn polynomial vector weight functions. Aggregates of these spacings components provide natural EQF analogues of the Neyman smooth statistics, an orthogonal decomposition of the Greenwood spacings statistic, and a discrete spacings analogue of the Anderson-Darling statistic. Asymptotic distribution theory is obtained under uniformity as well as fixed and local alternatives.

Results from Monte Carlo studies indicate adequacy of asymptotics for small samples and suggest a hybrid EDF-EQF quadratic statistic as an omnibus test for uniformity.

Subsampling Quantile Estimator Standard Errors With Applications

W. D. Kaigh and C. Cheng (1991)
Communications in Statistics, Part A, 20, 977-995.

Several smoothed quantile estimators recently have been proposed as alternatives to the conventional sample quantile, and previous studies have shown that the Harrell and Davis (1982), Kaigh and Lachenbruch (1982), and Kaigh and Cheng (1990) subsampling estimators usually improve efficiency of population quantile point estimation. Bootstrap and jackknife standard error formulas are developed here for all three subsampling quantile estimators. Applications to one-sample quantile confidence intervals and tests for equality of two population medians are illustrated with Monte Carlo results to indicate that the proposed methods provide improvements over existing procedures.

D. Participating Scientific Personnel

The following scientific personnel were supported by the project and contributed to the research effort. Only the principal investigator was supported throughout the three-year effort.

Dr. W. D. Kaigh (U. T. El Paso Professor of Mathematical Sciences): Principal Investigator

Meilin Yan (U. T. El Paso Graduate Student): Project Research Assistant

Ms. Yan received the Masters' of Science Degree in Statistics on December 15, 1991.

Alexander White (U. T. El Paso Graduate Student): Project Research Assistant

Mr. White received the Masters' of Science Degree in Statistics on August 30, 1992.

Thesis Title-- "EQF Component Decompositions With Applications"

Thesis Director-- Dr. W. D. Kaigh

(This thesis received the U. T. El Paso Outstanding Graduate Student Thesis Award.)

Maria Alejandra Sorto (U. T. El Paso Graduate Student): Project Research Assistant

Ms. Sorto received the Masters' of Science Degree in Statistics on August 30, 1992.

Thesis Title-- "Subsampling Quantile Estimators"

Thesis Director-- Dr. W. D. Kaigh

REPORT OF INVENTIONS

No inventions were developed under the project.

BIBLIOGRAPHY

Durbin, J. and Knott, M. (1972), "Components of Cramer-von Mises Statistics I," *Journal of the Royal Statistical Society, Ser. B*, 34, 290-307.

Kaigh, W. D. (1992), "EDF and EQF Orthogonal Component Decompositions," *Nonparametric Statistics*, 1, 313-334.

Kaigh, W. D. (1993), "Distribution-Free Two-Sample Tests Based on Rank Spacings," *Journal of the American Statistical Association*. In press.

Kaigh, W. D. and Cheng, C. (1991), "Subsampling Quantile Estimator Standard Errors With Applications," *Communications in Statistics, Part A-Theory and Methods*, 20, 977-995.

Kaigh, W. D. and Hosch, H. M. (1993), "Quartile Estimation Techniques: Explanation and Application to Data Analysis in Psychology." Submitted for publication.

Kaigh, W. D. and Sorto, M. A.. (1993), "Subsampling Quantile Estimator Majorization Inequalities," *Statistics and Probability Letters*. In press.

Kaigh, W. D. and White, A. K. (1993), "Total Time on Test Function Orthogonal Components and Tests of Exponentiality," *Proceedings of the 37th Conference on the Design of Experiments*. In press.

Parzen, E. (1979), "Nonparametric Statistical Data Modeling" (with comments), *Journal of the American Statistical Association*, 74, 105-121.

Rayner, J. C. W. and Best, D. J. (1989), *Smooth Tests of Goodness of Fit*, New York: Oxford University Press.

Shorack, G. R. and Wellner, J. A. (1986), *Empirical Processes With Applications to Statistics*, New York: John Wiley.

Sorto, M. A. (1992), "Subsampling Quantile Estimators," University of Texas at El Paso Masters' Thesis.

White, A. K. (1992), "EQF Component Decompositions With Applications," University of Texas at El Paso Masters' Thesis.